EXPERIMENTS IN SPEECH DRIVEN INFORMATION RETRIEVAL FOR SPANISH LANGUAGE

César González-Ferreras, Valentín Cardeñoso-Payo

Departamento de Informática
Universidad de Valladolid
{cesargf,valen}@infor.uva.es

ABSTRACT

The paper reports on the evaluation of a system that allows users to search information using spoken queries. The front end is a large vocabulary continuous speech recognizer which translates the query from speech to text and puts it through an information retrieval engine to retrieve the set of relevant documents for that query. The system is designed for Spanish language. The performance of the system was evaluated using the test suites of CLEF, which is an evaluation forum similar to TREC. 10 different speakers were recorded reading the queries. Results of different experiments are reported. Best results were obtained with a language model of 60,000 words and medium length queries: loss of precision of 23.79% (compared to using perfect transcription of the queries) with a WER of 18.4%. Overall, these results are encouraging and provide a solid basis for the feasibility of building speech driven information retrieval systems.

1. INTRODUCTION

Nowadays web contents are an important source of information. We use the web to access different kinds of information, both for leisure and business. There is also a growing interest in providing speech access to web contents. Using speech is natural for most of the people, and thus it can provide a more usable interaction. Mobile devices allow web access anytime and everywhere, and would also benefit from speech interaction, increasing their usability, which is limited because of their small displays.

Different approaches have been proposed to access web contents using speech. One of the most natural and effective is using speech as the input to an information retrieval engine. In fact, search is one of the most used ways to access the world wide web. Search can also help to deal with the limitations of the speech channel, which does not allow to send much information over it.

This paper reports on the evaluation of a system that provides spoken access to an information retrieval engine. The front end is a large vocabulary continuous speech recognizer (LVCSR) which translates the query from speech to text and puts it through an information retrieval (IR) engine to retrieve the set of relevant documents for that query. The system is designed for Spanish language.

The performance of the system was evaluated using the test suites of CLEF, which is an evaluation forum similar to TREC. Results of different experiments are reported. The impact of vocabulary size was tested first, comparing results of trigram language models (LM) with 20,000 and 60,000 words vocabulary sizes. Then, the influence of query length was measured, using medium length and short queries. Best results were obtained with LM of 60,000 words and medium length queries: loss of precision of 23.79% (compared to using perfect transcription of the queries) with a word error rate (WER) of 18.4%.

The structure of the paper is as follows: section 2 presents some related work; section 3 explains the system in detail; section 4 describes the experiments and shows the results; in section 5 we discuss about factors that affect system performance; section 6 presents conclusions and future work.

2. RELATED WORK

There has been little work on speech driven information retrieval. Some experiments have been carried out in English, Chinese and Japanese. All the experiments used a similar methodology: a standard IR test collection (designed to evaluate IR systems using text queries under standard and comparable conditions) is used; some speakers reading the queries are recorded; finally, system performance is evaluated and compared with results obtained using text queries.

First speech driven information retrieval experiments have been described in [1]. Dragon's LVCSR system with 20,000 and 30,000 words vocabulary and INQUERY probabilistic IR system were used. Two test collections were used: TIPSTER and Boston Globe. Experiments were reported using long, medium length and short queries. Results showed that increasing WER reduces precision and that long queries are more robust to recognition errors

This work has been partially supported by *Ministerio de Ciencia y Tencología* under project number TIC2003-08382-C05-03 and by *Consejería de Educación de la Junta de Castilla y León* under project number VA053A05.

than short queries. Based on these experiments, an interactive vocal information retrieval system was designed [2]. The system had the following components: a dialogue manager, a probabilistic IR system, a document summarization system and a document delivery system.

Experiments in Chinese have been presented in [3]. Two different speech recognizers were used: Microsoft SAPI 5.0 LVCSR (50,000 Chinese multi-character words) and HTK-based syllable speech recognition system. The IR system used the vector space model with tf-idf and pseudo relevance feedback. TREC5 and TREC6 databases were used as test collections. Retrieval performance on mobile devices with high-quality microphones (for example PDA) was satisfactory, although the performance over cellular phone was significantly worse.

Experiments in Japanese have been reported in [4]. Julius LVCSR with trigram LM of 20,000 and 60,000 words was used. The IR system was based on the probabilistic model. NTCIR-3 test collection was used. Results showed that using target document collection for language modeling and using bigger vocabulary size improved system performance. More experiments using the same test collection have been described in [5]. Julius and Spojus LVCSRs, using phone-based and syllable-based models were used. The IR system used the vectorial model and tf-idf. Different techniques for combining multiple LVCSRs using SVM learning were evaluated. Results showed an improvement both in speech recognition and retrieval accuracies. They also reported better results with a LM of larger vocabulary size.

3. SYSTEM OVERVIEW

The objective of the system is to retrieve all the documents relevant to a given spoken query. The architecture of the system is shown in figure 1. First, the user makes a query using speech. Then, the speech recognizer transcribes the spoken query into text. Finally, the information retrieval engine finds the list of documents relevant to that query. In the following sections we describe in detail both speech recognition and information retrieval systems.

3.1. Speech Recognition

For speech recognition we used SONIC, the University of Colorado large vocabulary continuous speech recognizer [6]. SONIC is based on continuous density hidden Markov model (CDHMM) technology and implements a two-pass search strategy using token-passing based recognition.

The triphone acoustic models were HMMs with associated gamma probability density functions to model state durations. Mel Frequency Cepstral Coefficients (MFCC) were used, with a standard 39-dimensional feature vector (12 MFCCs and normalized log energy along with the first and second order derivatives). We trained the acoustic models using Albayzin corpus [7] (13600 sentences read by 304 speakers).

Two trigram LMs with different vocabulary size were built: 20,000 and 60,000 words. The target document collection was used to train the language model, because this can result in an adaptation of the LVCSR to the given task and provides better system performance [4]. EFE94 document collection is composed of one year of newswire news (511 Mb). CMU-Cambridge statistical language modeling toolkit was used with Witten Bell discounting [8].

3.2. Information Retrieval

A modified version of an information retrieval engine developed for Spanish language was used [9]. It is based on the vector space model and term frequency-inverse document frequency (tf-idf) weighting scheme [10]. A stemming algorithm was used to reduce the dimensionality of the space and a stop word list to remove function words.

Given a query q, the similarity of that query with each document d_i in the document collection is calculated as follows:

$$sim(d_i, q) = \sum_{t_r \epsilon q} w_{r,i} \times w_{r,q} \tag{1}$$

$$w_{r,i} = (1 + log(tf_{r,i})) \times log\left(\frac{N}{df_r}\right) \tag{2}$$

$$w_{r,q} = log\left(\frac{N}{df_r}\right) \tag{3}$$

Where $tf_{r,i}$ represents the number of times a term t_r appears in the document d_i; df_r is the number of documents in the collection in which the term t_r appears; N is the number of documents in the collection.

4. EXPERIMENTS

In order to evaluate the performance of the system we made some experiments based on Cross-Language Evaluation Forum (CLEF) test-suites [11]. CLEF organizes evaluation campaigns in a similar way to TREC. Its aim is to develop an infrastructure for the testing and evaluation of information retrieval systems operating on European languages, in both monolingual and cross-language contexts. We have expanded CLEF 2001 Spanish monolingual IR test-suite to include spoken queries. In the following sections the experimental set-up is described and the results are presented.

4.1. Experimental Set-up

CLEF 2001 test suite includes a document collection, a set of topics and relevance judgements. The document collection has 215,738 documents of the year 1994 from EFE newswire agency (511 Mb). There are 49 topics (topic numbers 41-90) and each of them has three parts:

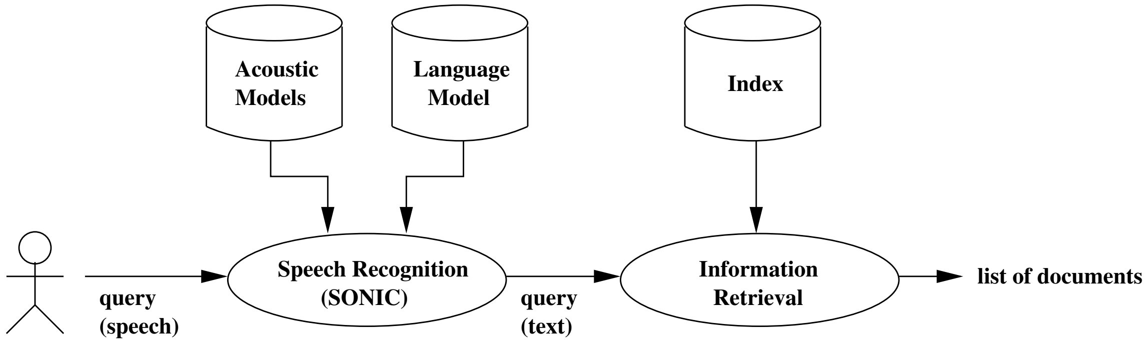

Figure 1. *Architecture of the system.*

```
<top>
<num>67</num>
<title>Ship Collisions</title>
<desc>Find information on the number
of people injured or killed in
collisions between ships.</desc>
<narr>Relevant documents report
information on the number of victims
(dead or injured) of collisions
between ships or maritime vessels of
all types.  Documents that speak of
victims without providing figures are
not relevant.</narr>
</top>
```

Figure 2. *Sample query (translated from Spanish).*

a brief title statement, a one-sentence description and a more complex narrative specifying the relevance assessment criteria (see figure 2). Relevance judgements determine the set of relevant documents for each topic, and were created using pooling techniques.

Given the topics, we tested our system using two different queries: (1) medium length queries (using description field, with mean length of 16 words, ranging from 5 to 33); (2) short queries (using title field, with mean length of 4 words, ranging from 1 to 8). We recorded queries of 10 different speakers (5 male and 5 female) using a headset microphone under office conditions, at 16 bit resolution and 16 kHz sampling frequency.

4.2. Results

First, spoken queries were processed by the speech recognizer and word error rate was calculated. Then, the best hypothesis was used by the information retrieval engine. The same methodology of CLEF was used to evaluate the results [11]. For each query, the 1000 most relevant documents (sorted by relevance) were retrieved, and using the relevance judgements mean average precision (MAP) was calculated.

Results are shown in tables 1 and 2. In the first table a LM of 20,000 words was used, with an OOV rate of 7.07%. In the second table a LM of 60,000 words was used, with an OOV rate of 2.02%. In each table we compare results using medium length queries (description) and short queries (title). For each speaker, we report the word error rate (WER), the mean average precision (MAP), the loss of MAP compared with using the original text queries (Ploss), and the correlation between WER and loss of MAP (corr).

5. DISCUSSION

We analyzed the impact of LM vocabulary size. Better results were obtained using a LM of 60,000 words, because there were fewer speech recognition errors. Speech driven information retrieval is a task with an open vocabulary, and better results will be obtained with larger vocabulary LM, because of its better vocabulary coverage.

We studied the impact of query length. Medium length queries had better results of precision than short queries, both in text and speech. Moreover, medium length spoken queries had fewer loss of precision than short spoken queries, and thus they seem to be more robust to speech recognition errors.

A correlation between WER and loss of precision was detected. The correlation was stronger in short queries than in medium length queries. The reason is that most of the words in short queries are keywords, so any recognition error has big impact in retrieval results. However, in medium length queries the recognition errors might happen in function words and do not degrade precision.

We analyzed the results of each individual query. Most of the queries had small loss of precision while some of the queries had high loss of precision. It means that in general queries did well, but there were some that did badly. For the experiment with a LM of 60,000 words and medium length queries, 70.61% of the queries had a loss of precision of less than 10% and 9.39% of the queries had a loss of precision between 10% and 30%. However, 11.63% of the queries had a loss of precision more than 90%.

	description				title			
	WER	**MAP**	**Ploss**	**corr**	**WER**	**MAP**	**Ploss**	**corr**
Text		0.4466				0.4373		
Speaker1	24.9%	0.2887	35.36%	0.42	31.4%	0.2540	41.92%	0.76
Speaker2	17.4%	0.3134	29.83%	0.28	23.2%	0.2980	31.85%	0.73
Speaker3	22.7%	0.2980	33.27%	0.22	18.8%	0.3094	29.25%	0.66
Speaker4	23.5%	0.2921	34.59%	0.41	26.1%	0.2795	36.09%	0.59
Speaker5	28.2%	0.3110	30.36%	0.47	23.2%	0.3017	31.01%	0.57
Speaker6	33.6%	0.2404	46.17%	0.50	27.5%	0.2755	37.00%	0.63
Speaker7	15.8%	0.3282	26.51%	0.44	21.3%	0.2928	33.04%	0.63
Speaker8	23.1%	0.3005	32.71%	0.51	26.1%	0.2926	33.09%	0.71
Speaker9	24.4%	0.2926	34.48%	0.36	21.7%	0.2713	37.96%	0.74
Speaker10	27.8%	0.3271	26.76%	0.43	26.1%	0.2911	33.43%	0.75
Mean	24.1%	0.2992	33.00%	0.41	24.5%	0.2866	34.46%	0.66

Table 1. *Results using a LM of 20,000 words for medium length queries (description) and short queries (title). (WER: word error rate; MAP: mean average precision; Ploss: loss of MAP compared with text queries; corr: correlation between WER and Ploss).*

	description				title			
	WER	**MAP**	**Ploss**	**corr**	**WER**	**MAP**	**Ploss**	**corr**
Text		0.4466				0.4373		
Speaker1	18.5%	0.3282	26.51%	0.62	25.1%	0.2938	32.82%	0.79
Speaker2	10.7%	0.3618	18.99%	0.60	12.6%	0.3586	18.00%	0.82
Speaker3	16.5%	0.3514	21.32%	0.29	11.1%	0.3513	19.67%	0.74
Speaker4	17.1%	0.3269	26.80%	0.47	14.5%	0.3362	23.12%	0.85
Speaker5	22.5%	0.3401	23.85%	0.47	15.9%	0.3305	24.42%	0.82
Speaker6	27.7%	0.2950	33.95%	0.56	20.3%	0.3239	25.93%	0.67
Speaker7	8.9%	0.3587	19.68%	0.32	14.5%	0.3176	27.37%	0.62
Speaker8	18.5%	0.3359	24.79%	0.63	15.5%	0.3320	24.08%	0.83
Speaker9	20.9%	0.3236	27.54%	0.41	13.5%	0.3297	24.61%	0.90
Speaker10	22.7%	0.3818	14.51%	0.50	18.4%	0.3305	24.42%	0.75
Mean	18.4%	0.3403	23.79%	0.48	16.1%	0.3304	24.44%	0.71

Table 2. *Results using a LM of 60,000 words for medium length queries (description) and short queries (title). (WER: word error rate; MAP: mean average precision; Ploss: loss of MAP compared with text queries; corr: correlation between WER and Ploss).*

6. CONCLUSIONS

In this paper we have described a system that allows users to retrieve information using spoken queries. We made different experiments in order to evaluate the system performance using a standard IR test suite. The impact of vocabulary size and query length were analyzed. Best results were obtained with a LM of 60,000 words and medium length queries: loss of precision of 23.79% (compared to using perfect transcription of the queries) with a WER of 18.4%. We also reported correlation between WER and loss of precision. Overall, these results are encouraging and provide a solid basis for the feasibility of building speech driven information retrieval systems.

As future work, we will focus on queries with high precision loss. We plan to study the factors that explain their performance decrease and to develop a method to avoid them. Another area of interest is related to the construction of spoken dialog systems, where user interaction could provide valuable feedback to improve the retrieval process.

7. ACKNOWLEDGEMENTS

The authors want to thank Joaquín Adiego and Pablo de la Fuente for their help in the information retrieval part of the system.

8. REFERENCES

[1] J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, and S. W. Kuo, "Experiments in Spoken Queries for Document Retrieval," in *European Conference on Speech Communication and Technology (Eurospeech)*, 1997.

[2] F. Crestani, "Vocal Access to a Newspaper Archive: Assessing the Limitations of Current Voice Information Access Technology," *Journal of Intelligent Information Systems*, vol. 20, no. 2, pp. 161–180, 2003.

[3] E. Chang, F. Seide, H. Meng, Z. Chen, Y. Shi, and Y. Li, "A System for Spoken Query Information Retrieval on Mobile Devices," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 8, pp. 531–541, November 2002.

[4] A. Fujii and K. Itou, "Building a Test Collection for Speech-Driven Web Retrieval," in *European Conference on Speech Communication and Technology (Eurospeech)*, 2003.

[5] M. Matsushita, H. Nishizaki, S. Nakagawa, and T. Utsuro, "Keyword Recognition and Extraction by Multiple-LVCSRs with 60,000 Words in Speech-driven WEB Retrieval Task," in *International Conference on Spoken Language Processing (ICSLP)*, 2004.

[6] B. Pellom and K. Hacioglu, "Recent Improvements in the CU SONIC ASR System for Noisy Speech: The SPINE Task," in *International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2003.

[7] A. Moreno, D. Poch, A. Bonafonte, E. Lleida, J. Llisterri, J.B. Mariño, and C. Nadeu, "ALBAYZIN Speech Database: Design of the Phonetic Corpus," in *European Conference on Speech Communication and Technology (Eurospeech)*, 1993.

[8] P. Clarkson and R. Rosenfeld, "Statistical Language Modeling Using The CMU-Cambridge Toolkit," in *European Conference on Speech Communication and Technology (Eurospeech)*, 1997.

[9] J. Adiego, P. Fuente, J. Vegas, and M. Á. Villarroel, "System for Compressing and Retrieving Structured Documents," *UPGRADE*, vol. 3, no. 3, pp. 62–69, June 2002.

[10] G. Salton, A. Wong, and C. S. Yang, "A Vector Space Model for Automatic Indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, November 1975.

[11] M. Braschler and C. Peters, "CLEF Methodology and Metrics," in *Workshop of the Cross-Language Evaluation Forum (CLEF)*, 2001.